学校编码：10384
学号：19820061151813

分类号_____密级_____
UDC _____

厦 门 大 学

硕 士 学 位 论 文

核磁共振代谢组学数据处理新方法及应用

New methods of NMR-based metabonomics data processing and its application

徐乐

指导教师姓名：董继扬 副教授

专 业 名 称：无线电物理

论文提交日期：2009 年 5 月

论文答辩时间：2009 年 6 月

学位授予日期：2009 年 月

答辩委员会主席：_____

评 阅 人：_____

2009 年 6 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文而产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（　），在　　年解密后适用本授权书。

2、不保密（　）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

目　录

CONTENTS

摘　要

代谢组学是关于定量描述生物内源性代谢物质的整体及其对内因和外因变化应答规律的科学，是系统生物学的有机组成部分。以核磁共振(Nuclear Magnetic Resonance，NMR)为主要分析手段的代谢组学通常称为NMR代谢组学，已经广泛地应用到了包括药物研发，分子生理学，分子病理学，基因功能组学，营养学，环境科学等重要领域。

NMR代谢组学的数据具有非线性、高维性、小样本性等特点，开发符合这些特性、且具有一定普适性的数据分析方法是代谢组学发展的关键。本文从数据预处理和数据分析两个方面入手，提出基于数据统计差异性的自适应分段积分方法和基于非负矩阵分解算法的数据分析方法。本文的主要内容如下：

一、简要综述了目前代谢组学研究中常用的模式分析方法，讨论了代谢组学数据分析的发展趋势。

二、提出了基于统计差异性的自适应分段积分数据预处理方法。提出描述数据统计差异性的函数，并根据变量的统计差异性自适应地选择积分间隔，实现数据矩阵的自适应分段积分。采用模拟数据集和素食研究的真实数据集对算法有效性进行验证，结果表明，自适应分段积分有助于提高样品分类与特征代谢物寻找的准确性。

三、将非负矩阵分解(NMF)算法引入 NMR 代谢组学模式分析中，分析 II 型糖尿病病人与健康人的血液及尿液样品，得到与 II 型糖尿病相关的一些特征代谢物。通过与 PCA 分析结果的比较，显示了 NMF 算法基于数据非负性和局部表示的思想更适于小浓度标记代谢物的检测。

研究表明，基于数据统计差异性的自适应分段积分方法和基于非负矩阵分解算法的模式识别方法能够得到更可靠的模式识别分析结果，使寻找到的特征代谢物更具有生物学意义。

关键词：代谢组学；分段积分；数据分析

Abstract

Metabonomics is the branch of science concerned with the quantitative understandings of the metabolite complement of integrated living systems and its dynamic responses to the changes of both endogenous factors and exogenous factors. NMR-based metabonomics which uses nuclear magnetic resonance as the main analysis technology has been widely used in many fields, including drug research, molecular physiology, molecular pathology, genomics, nutrition, environment science.

Due to the nonlinear, high dimensionality and small sample size characteristics of metabonomics data, it is a key problem to development new data analysis methods in line with these characteristics and with certain universality. This paper proceeded with data pretreatment and data analysis. In our work, a novel adaptive binning method based on statistical discrepancy and a multivariate statistical analysis method based on non-negative matrix factorization (NMF) were introduced. The main results are summarized as follows:

First, with substantive literature，make a brief overview of common pattern analysis methods, and prospect the development trend of metabonomics data analysis.

Second, a novel adaptive binning method based on statistical discrepancy was proposed for data pretreatment. A function is constructed to describe the statistical discrepancy of metabonomics data. Then the data matrixes are integrated with the integral interval designed adaptively based on the statistical discrepancy of variables. Both simulated NMR data and experimental spectra from dietary intervention individuals were employed to validate the performance of the adaptive binning. It was showed that the accuracy of sample classification and characteristic biomarkers identification can be improved effectively by the proposed binning method.

Third, non-negative matrix factorization (NMF) was applied to the NMR-based metabonomics pattern analysis. Detail comparisons were made between NMF and the most conventional method principal component analysis (PCA) by employing the two

methods to discriminate the urine and serum spectra of diabetes II patients from healthy controls and identify the potential biomarkers. It was proved that the special advantage of NMF such as the non-negative constraints and the part-based representation are more feasible for detecting small concentration biomarkers.

It could be concluded that，the adaptive binning method based on statistical discrepancy and the multivariate statistical analysis method based on NMF are effective tools for pattern analysis and characteristic biomarkers identification in metabonomics research.

Keywords: metabonomics; adaptive binning; data analysis

第一章 绪论

1.1 基于NMR的代谢组学概述

随着人类基因组测序工作的完成，基因功能的研究逐渐成为热点，随之出现了一系列的“组学”研究，包括研究转录过程的转录组学（transcriptomics）、研究某个生物体系中所有蛋白质及其功能的蛋白质组学（proteomics）及研究代谢产物的变化及代谢途径的代谢组学（metabolomics/metabonomics）。

代谢组学（metabonomics/metabolomics）是继基因组学、转录组学、蛋白质组学后系统生物学的另一重要研究领域，是一种定量考察生命系统由于受到外源性刺激或者基因修饰而产生的与时间相关的多参数代谢应答的研究方法[1,2]。它利用体液和组织等生物样品进行疾病诊断，是实现快速、非侵入性疾病诊断的有效途径。代谢组学的概念最早来源于代谢轮廓分析[3]。Nicholson研究小组于1999年提出了代谢组学的概念[4]，并在疾病诊断、药物筛选等方面做了大量的卓有成效的工作。Fiehn[5]于1997年提出了metabolomics的概念，第一次把代谢产物和生物基因的功能联系起来，之后很多植物化学家开展了植物代谢组学的研究，使得代谢组学得到了极大的充实，同时也形成了当前代谢组学的两大主流领域：metabolomics和metabonomics。作为崭新的方法学，代谢组学已成为国际上疾病与健康研究的一个重要热点。目前，代谢组学已经在疾病诊断、药物作用机制研究和安全性评价等方面表现出重要的理论意义和应用价值[6,7]。

根据研究目的和对象的不同，生物体代谢产物分析可分为4个层次[8]：（1）代谢物靶标分析（MTA：Metabolite Target Analysis）。某个或几个特定组分的分析。（2）代谢轮廓分析（MP：Metabolic Profiling）。少数预设的一些代谢产物的定量分析，如某一类结构、性质相关的化合物（氨基酸、有机酸顺二醇类）或某一代谢途径的所有中间产物或多条代谢途径的标志性组分。（3）代谢组学（Metabolomics/Metabonomics）。限定条件下对特定生物样品中所有代谢组分的定性和定量分析。（4）代谢物指纹分析（MFP：Metabolic Finger Printing）。不分离鉴定具体单一组分，而是对样品进行快速分类（如表型的快速鉴定）。目前代谢组学研究主要集中在第3层次。